In Natur und Freiheit: Akten des XII. Internationalen Kant-Kongresses (2015), Eds. Violetta L. Waibel, Margit Ruffing, and David Wagner. Berlin/Boston: De Gruyter, 2018. Pages 2037-2045.

Adrian M. S. Piper

The Logic of Kant's Categorical "Imperative"

In Kant's moral philosophy, the imperative is perched precariously between two realms governed by the indicative form of speech. In the *Critique of Pure Reason*, it reminds us of an intelligible, rational realm beyond that governed by descriptive causal explanation. But in the *Groundwork of the Metaphysics of Morals*, it reminds us of the sensible pull of causality that frequently leads us to violate its intelligible principles. Correspondingly, Kant approaches the moral law from two directions in the *Groundwork of the Metaphysic of Morals*. When he is explaining moral motivation, he describes and refers to the moral law as an imperative, because this is the form it takes for causally enmeshed human beings. But when he is analyzing its rational formal structure and situating it within his broader analysis of reason, he formulates the moral law in the indicative mood, because this is the form it takes for perfectly rational beings.

The vast majority of Kant's actual formulations of the moral law in the *Groundwork* are not in the imperative. Of 47 formulations to be found in the text, only four are in the strict imperative. Of those four, only the first[1] receives extended analysis. Of the remaining 43, 31 are expressed in the indicative mood. So two-thirds of Kant's formulations of the moral law in the *Groundwork* are descriptive rather than prescriptive. These serve to buttress his repeated reminder that for agents as members of the intelligible world, the "I ought" becomes an "I will"[2]. In the *Groundwork*, Kant usually addresses us as members of that world.

Kant's overriding reliance on descriptive, categorical indicative formulations of the moral law is consistent with the theory of reason he offers in the *Critique of Pure Reason*. There, reason generates transcendent principles, concepts and theories also formulated in the indicative mood. These describe an ideal reality that regulates our empirical thoughts and actions. Kant tries to show that these transcendent ideas are "in the first place, simply categories extended to the unconditioned"[3], i.e. structured and engendered by the logical categories listed in the

[1] Kant: GMS, AA 04: 402.09 – 11.

[2] Kant: GMS, AA 04: 455.08 – 12; cf. 414.01 – 13, 439.36 – 40, 449.20 – 28 and 454.09 – 16.

[3] Kant: KrV, A 409.20 – 24/B 436.15 – 19.

Adrian M. S. Piper, APRA Foundation Berlin, contact@adrianpiper.com

https://doi.org/10.1515/9783110467888-193

first *Critique*'s Table of Judgments, and given incomplete content by intuition in the Table of Categories.

Kant regards the moral law as a principle of action similarly engendered by the logical categories of the Table of Judgments, and similarly descriptive of an ideal reality that regulates our empirical thoughts and actions. In the first *Critique*, he identifies action as a "pure but derivative," *a priori* "predicable" of the category of causality.[4] The concept of action is "pure" in its transcendental and *a priori* status, as a necessary precondition of our experience of our own and others' agency. It is "derivative" in that it is an instance or particular kind of causality required for experience. Finally, the concept of action is a "predicable" in that it is the concept of a particular kind of effect that is to be predicated of its particular kind of cause, namely the agent.

Ascribing transcendental status to the concept of action thereby distinguishes it from empirical concepts, which contingently instantiate the pure concepts of the understanding. By contrast, the concept of action is a universally valid precondition of unified experience under any circumstances. But if the logical structure of the transcendental categories engenders the transcendent Ideas of Reason in general, as Kant claims, then in particular, Kant's ascription of transcendental status to the concept of action engenders the transcendent Idea of perfectly rational action that both expresses "the capacity to act from freedom"[5] and also thereby anchors the *Groundwork*. It is this idea, both descriptive of an ideal reality and also regulative of our empirical conduct, which Kant's many categorical indicative formulations of the moral law attempt to capture.

All categorical indicative statements ascribe a predicate to a subject. If a cause in general is a subject of which we predicate its effect, then even the category of causality in the Table of Categories, and therefore the corresponding hypothetical and disjunctive relational forms in the Table of Judgments, ultimately have the same logical subject-predicate form that we find in Kant's remarks about action at KrV, A 80/B 108.16. The subject-predicate judgment form expresses in natural language the logical relationship Fa between any subject a and a property F that can be predicated of it; and therefore between any cause and its effect.

In particular, it expresses the logical relationship between an agent w and her action A in the categorical indicative expression Aw, in which action A is a property we ascribe to agent w. Kant's identification of an action A as an effect to be ascribed to an agent w as its cause,

[4] Kant: KrV A 80/B 108.16.

[5] Kant: KrV A 450.07 f./B 478.07 f.

(1) *Aw*

expresses the categorical indicative statement, "*w* does *A*." Then treat *Aw* as itself the object of *w*'s intention *P*, such that

(2) *Pw(Aw)*.

(2) expresses the categorical indicative statement, "*w* wills [or intends] to do *A*," i.e. it denotes agent *w*'s intention. Here the expression *Pw* is treated not as an operator, but rather as itself a predicate of whatever lies between the outermost brackets that follow it. Because *P* predicates an intentional state of *w*, the same opacity constraints on inference apply nevertheless.

This description of an action as the agent's intended end is Kant's notion of a maxim. Kant characterizes a maxim as a categorical indicative by implication, when he first calls it a "subjective principle of action," and next contrasts it with an objective principle or law "on which [one] *ought to act* – that is, an imperative"[6]. So whereas an objective principle of action is an imperative, a maxim is not. Syntactically, a maxim is a first-person action description. Kant's more careful formulations of the maxims in the examples he discusses in the *Groundwork* give them a three-fold structure:

(3) M_1: Out of respect for the moral law [= ground],
M_2: I will pay my bills in a timely manner [= will],
M_3: in order to discharge my financial obligations [= end].

In (3), the "out of" locution M_1 identifies the backward-looking motive, or what Kant calls the *ground*. The "I will" locution M_2 identifies the intention, or what Kant calls the *will*. And the "in order to" locution M_3 identifies the forward-looking goal of the action, or what Kant calls the *end*. But it is the intentional subject-predicate structure *Pw(Aw)* of M_2 that identifies the core maxim.

The self-legislation procedure Kant proposes in the *Groundwork* is directed at maxims that express and enact respectful attention to the deliverances of moral principle. All of his examples require choosing between a maxim of selfish inclinational behavior and a maxim of ethically principled conduct. In each, Kant believes that a positive answer to his question, "Can you also will, that your maxim become a universal law?"[7] identifies the right course of action and motivates choosing the ethical route over the selfish one. Kant's insistence that we choose that act whose maxim we can will as a universal law is, of course, the requirement usually identified as Kant's categorical "imperative." The suggestion here

[6] Kant: GMS, AA 04: 421 n.

[7] Kant: GMS, AA 04: 403.26 f.

is that this requirement is more accurately described as Kant's categorical *indicative*. Using the notation developed so far, this requirement can be formulated symbolically as

(4) $Pw(Aw) \rightarrow \Diamond Pw[(\forall x)(Ax)]$ *Kant's Categorical Indicative*

(4) says that agent *w* wills to do *A* only if it is possible for *w* to will (i.e. only if *w* could will) that everyone does *A*. (4) is cast in the indicative mood that describes the behavior of a perfectly rational being whom we are moved by respect for this very descriptive principle to emulate.

(4) is plausible only if the expression "$(\forall x)(Ax)$" is interpreted distributively rather than collectively. In quantificational terms, "$(\forall x)(Ax)$" is to be understood not conjunctively, as stipulating that one must be able to will that everyone together does *A*; but rather disjunctively, to stipulate that one must be able to will that any individual does *A* under the relevant circumstances. So, for example, if

(5) $Pi(Ai)$

symbolizes the maxim, "I will pay my monthly bills in a timely manner," (4) requires not that everyone conjointly pay their monthly bills in the same timely manner; but rather merely a settled universal practice that each individual pays her monthly bills when they come due. The distributive interpretation would be the natural one, were (4) a law of nature in the sense required by Kant's formulation of the moral law at GMS, AA 04: 421.21–23.

Kant argues that I must be able to will a permissible maxim as a universal law; and that I can do that only if there is no contradiction in my will when I do so. He clearly explains what he means by a contradiction in the will when he describes our futile attempts to will the universalization of a selfish maxim:

> (6) [W]e find that we do not really will such that our maxim should become a universal law, [...] but rather the opposite is itself in reality to remain universally a law.[8]

Here he describes the internally conflicted state in which we both try to will the universalization of the derelict maxim, and at the same time "do not really will" it. He then observes that if we

[8] Kant: GMS, AA 04: 424.19–22.

> (7) consider everything from one and the same standpoint, namely reason, we would come across a contradiction in our own will, namely that a particular principle should be objectively necessary as universal law, and yet subjectively not be universally valid, but rather should allow exceptions.[9]

Thus there is a contradiction in my will if I both will that everyone does *A*, and also "do not really will" this; that is, if I both do and do not will its universal validity. We can represent this self-contradictory condition symbolically as

(8) $\Diamond Pw[(\forall x)(Ax)] \rightarrow \Box{\sim}\{[Pw[(\forall x)(Ax)] \,.\, {\sim}Pw[(\forall x)(Ax)]\}$

Kant's Contradiction in the Will Test

(8) states that *w* can will that everyone does *A* only if it is impossible that *w* both wills this and also does not will this.

(8) is a tautology of the form, *P* only if not both *P* and not-*P*. That (8) is a tautology is a good thing. It shows that there is a close, recognizable and logically necessary connection between Kant's universalizability criterion of maxim validity and the conditions under which it is violated. It is violated just in case the universalization of the maxim functions as one of the two expressions in an ordinary logical contradiction. So Kant's universalizability requirement is in essence one of logical consistency. It is an application from the first *Critique* of his Highest Principle of All Analytic Judgments, namely the principle of noncontradiction, to the special case of action.

Transitivity on (4) and (8) then yields the implication that I can be said to properly will an action only if willing its universalization includes no such self-contradiction:

(9) $Pw(Aw) \rightarrow \Box{\sim}\{[\, Pw[(\forall x)(Ax)] \,.\, {\sim}Pw[(\forall x)(Ax)]]\}$ (4), (8)

(9) is another nontrivial tautology. It says that an agent *w* wills act *A* only if his will that everyone does *A* cannot contradict itself. (9) captures the substantive insight that a perfectly rational being who naturally and spontaneously acts in accordance with the moral law wills that action wholeheartedly, independently of any conflicting internal states. Whatever these may be, they by definition do not divert the agent into pseudorational dithering, regret, or self-contradiction. Delinquent or countervailing inclinations may be powerful without sabotaging the agent's consistent formulation of and steadfast commitment to her rational intention. Against the force of *perfectly* rational intention, any such inclinations are in any case irrelevant.

[9] Kant: GMS, AA 04: 424.25–30.

Let the argument consisting of (4), (8) and (9) constitute Kant's analysis of *rational resolve* as that brand of resolve, willing or intention that is universalizable without self-contradiction:

(10) Premise 1: $Pw(Aw) \rightarrow \Diamond Pw[(\forall x)(Ax)]$ (4)
Premise 2: $\Diamond Pw[(\forall x)(Ax)] \rightarrow \Box{\sim}\{[Pw[(\forall x)(Ax)] \,.\, {\sim}Pw[(\forall x)(Ax)]\}$ (8)
Conclusion: $\therefore Pw(Aw) \rightarrow \Box{\sim}\{[Pw[(\forall x)(Ax)] \,.\, {\sim}Pw[(\forall x)(Ax)]\}$ (9)

Now Kant's ambitious attempt to derive substantive morality from reason is not successful. So his analysis of rational resolve does not *imply* the morally right choice in any of his four examples. But it does *apply* to all of them. Consider his maxim instantiating false promising: "Whenever I believe myself short of money, I will borrow money and promise to pay it back, although I know this will never happen."[10] In accordance with $(3.M_2)$ above, condense this into

(11) $Pi(Bi)$,

where *Bi* is short for "I will make bad loans." Kant's more general argument against false promising is that

(12) I see immediately that my maxim could never be valid as an internally consistent universal law of nature, but rather must necessarily contradict itself. For the universality of a law that everyone who believes himself to be in need could promise whatever he wants with the intention not to keep it would make promising and whatever its point impossible.[11]

(12) applies to the particular instantiation in maxim (11). If I try to universalize it, I get

(13) $(\forall x)(Bx)$

i.e. everyone will make bad loans. But then everyone is not making loans at all, but rather defrauding their victims. Everyone cannot make bad loans without negating the very concept of a loan, which inherently involves repayment. So (13) implies its own negation:

(14) $(\forall x)(Bx) \rightarrow {\sim}[(\forall x)(Bx)]$.

[10] Kant: GMS, AA 04: 422.26–28.
[11] Kant: GMS, AA 04: 422.35–44.

(14) is not a formal contradiction. But it does show that I cannot consistently formulate the maxim that everyone will make bad loans (or, for that matter, bad promises of any kind) without thereby negating that maxim:

(15) $\{(\forall x)(Bx)] \rightarrow {\sim}[(\forall x)[(Bx)]\} \rightarrow {\sim}\Diamond Pi[(\forall x)(Bx)].$ (14)

And this forecloses its candidacy as "an internally consistent law of nature." Modus ponens on (14) and (15) yield

(16) ${\sim}\Diamond Pi[(\forall x)(Bx)],$ (14), (15)

which violates Premise (1) of Kant's analysis of rational resolve ((10), above). Hence maxim (11) does as well.

Generalizing now from *Pi* to any *Pw* and from bad promising to any action *Aw*, transposition and substitution on (15) yield

(17) $\Diamond Pw[(\forall x)(Ax)] \rightarrow {\sim}\{(\forall x)(Ax) \rightarrow {\sim}[(\forall x)(Ax)]\}$ (15)

Kant's Contradiction in Conception Test

(17) states that *w* can will that everyone does *A* only if everyone's doing *A* does not imply not everyone's doing *A*. (15) and (17) apply to false promising in general, as well as to any substitution instance of it.

But it is the conclusion of Kant's analysis ((9) above) that exposes the underlying contradiction in the *will* of the false promisor who attempts to conceptualize this incoherent idea: I both will to reap the advantage of making bad loans; and also "do not really will" the universalization of this intention, precisely because this "would make promising and whatever its point impossible:"

(18) $Pi(Bi) \rightarrow$ (9), (14)
${\sim}\Box{\sim}\{Pi[(\forall x)(Bx) \rightarrow {\sim}(\forall x)(Bx)] \;.\; {\sim}[Pi[(\forall x)(Bx) \rightarrow {\sim}(\forall x)(Bx)]]\}$

(19) $Pi(Bi) \rightarrow$ (18)
$\Diamond\{Pi[(\forall x)(Bx) \rightarrow {\sim}(\forall x)(Bx)] \;.\; {\sim}[Pi[(\forall x)(Bx) \rightarrow {\sim}(\forall x)(Bx)]]\}$

(18) and (19) attempt unsuccessfully to instantiate (8). (19) says that if I will to make bad loans, then I can will both that everyone makes bad loans, such that this implies not everyone making bad loans, and also not will this. However, the consequent of (19) is false, because logically contradictory. Since the antecedent is true, (19) itself is false. Kant explains the psychology of these self-contradictory mental gymnastics in passages (6) and (7) above: I will the incoherent principle of making of bad loans to hold for my own case, but thereby the negation of this principle to hold for everyone else's, because my intention to violate the principle myself can only be effective if others conform to it: The false promisor is a free rider.

So far I have exposed some of the logical substructure of Kant's categorical "imperative." I have shown how this predicate logic substructure is consistent with and can be integrated into Kant's comprehensive account of reason in the first *Critique*. I have defended its ability to evaluate an agent's decision as to whether or not to indulge the selfish inclination at the expense of the principled rational intention. However, that a maxim satisfies the criteria of rational resolve enumerated in (10) would at best identify its act as rational. This alone does not show that it is the best action to perform all things considered, because there may be several such alternatives that are both rational according to (10), and also mutually exclusive. Among those that do satisfy the rationality criteria expressed in (10), the categorical "imperative" procedure is silent on which one of them I should perform. So, for example, among the following three alternatives

(20) *Ai:* I will pay my monthly bills in a timely manner;

(21) *Ci:* I will pay off my student loan on schedule; and

(22) *Di:* I will wire the down payment on my new home mortgage to the bank by the designated deadline,

all three actions *Ai, Ci* and *Di* are equally rational and equally ethically permissible. But if my finances are limited and the deadlines for all three happen to coincide, then (20), (21) and (22) are mutually exclusive. Kant's self-legislation procedure does not enable us to choose among mutually incompatible actions all of whose maxims qualify as rational resolves.

Then *a fortiori*, Kant's procedure does not enable us to choose among universalizable act-descriptions those actions that are ethically permissible over those that are not. Suppose, for example, that I am undecided among the following alternatives:

(23) *Ai:* I will pay my monthly bills in a timely manner;

(24) *Ei:* I will day-trade my monthly salary in a high-frequency series of high-yield, high-risk equity investments; and

(25) *Fi:* I will divert my monthly salary into financing the purchase of black market artwork stolen from major museum collections.

Again the maxims of all three actions are, in fact, universalizable: I can will without internal contradiction that every salaried employee day-trade his monthly salary, as long as I get to day-trade mine (*Ei*). I can also consistently will that every such employee divert that salary into black market artwork (*Fi*). And again suppose that my finances enable me to act on only one of these alternatives. Just as before, each one of those alternatives counts as an object of choice. In order to decide what to do, I must choose among actions *Ai, Ei* and *Fi*.

An intuitively plausible ranking of these three alternatives would be

(26) *{Ai, Ei, Fi}*.

(26) assumes provisionally that the ethical act is preferred to the risky but potentially lucrative act; and that, in turn, preferred to the unethical act. But by itself, (26) expresses a merely subjective second-order preference for a particular ranking of first-order preferences. That is too *willkürlich* for Kant. The sorest point of dispute among Kantian and Humean action theorists is the question of whether Hume's provocation in the *Treatise* is true, that reason really is nothing but the "slave of the passions, and [...] '[t]is as little contrary to reason to prefer even my own acknowledg'd lesser good to my greater, and have a more ardent affection for the former than the latter".[12]

Can a preference ranking such as (26) be rationally justified in the broader sense that Kant intended his analysis of rational resolve to provide? In essence this question expresses the demand for a terminating, value-neutral criterion of rational final ends that caps the potentially infinite regress of higher and higher orders of preference over equally arbitrary subjective preference rankings. In fact there are ample resources elsewhere in Kant's theory of reason, particularly when buttressed by those of first-order predicate logic, which are uniquely suited to meet this demand. But that exploration must await another occasion.

[12] Hume, David: *A Treatise of Human Nature*. Ed. L. A. Selby-Bigge. Oxford 1968, T 415 f.